12 *De Novo* Structure Prediction: Methods and Applications

Richard Bonneau

1 Introduction

1.1 Scope of this Review and Definition of *De Novo* Structure Prediction

This review will focus on the questions: (i) what are the features common to methods that represent the current state of the art in *de novo* structure prediction and (ii) how can these methods benefit biologists whose primary aim is a systems-wide description of a given organism or system of organisms. The role and capabilities of *de novo* structure prediction as well as the relationship of *de novo* structure prediction to other sequence and structure-based methods is far from simple. The literature on this subject is rapidly evolving; for balance in coverage and opinion the reader is also referred to recent reviews of *de novo* structure prediction methods [11,27,32,39,50].

Many methods that are today referred to as *de novo* have alternately or previously been referred to *ab initio* or "new folds" methods. For the purpose of this review I will classify a method as *de novo* structure prediction if that method does not rely on homology between the query sequence and a sequence in the Protein Data Bank (PDB) to create a template for structure prediction. *De novo* methods, by this definition, are forced to consider much larger conformational landscapes than fold recognition and comparative modeling techniques that limit the exploration of conformational space to those regions close to the initial structural template or templates.

Another common pedagogical distinction between structure prediction methods has been the distinction between methods based on statistical principles, on the one hand, and physical or first principles, on the other hand. I will not discuss this distinction here at great length except for noting that one of the shortcomings of this artificial division is that most effective structure prediction methodologies are in fact a combination of these two camps. For example, several methods that are described as based on physical or first-principles employ energy functions and parameters that are statistical approximations of data (e.g. the Lennard–Jones representation of van der

Bioinformatics - From Genomes to Therapies Vol. 1. Edited by Thomas Lengauer
Copyright © 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-31278-8

Waals forces is often thought of as a physical potential, but is a heuristic fit to data). Most current successful *de novo* structure prediction methods fall into the statistics camp. A more useful distinction may be the distinction between reduced complexity models and models that use atomic detail. Throughout this chapter I will discuss low-resolution (models containing drastic reductions in complexity such as unified atoms and centroid representations of side-chain atoms) and high-resolution methods (methods that represent protein and sometimes solvent in full atomic detail) focusing on this practical classification/division of methods in favor of distinctions based on a given method's derivation or parameterization.

1.2 The Role of Structure Prediction in Biology

What is the main application of structure prediction to biology? At present this is an open question that will take many years to develop, as the answer relies on the relative rate of progress in several fields. In short, I will argue that the main current application of structure prediction in biology lies in understanding protein function. Structure predictions can offer meaningful biological insights at several functional levels depending on the method used to generate the structure prediction, the expected resolution and the comprehensiveness or scale on which predictions are available for a given system.

At the highest levels of detail/accuracy (comparative modeling) there are several similarities between the uses of experimental and computational/predicted protein structure and the types of functional information that can be extracted from models generated by both methods [4]. For example, experimentally determined structures and structures resulting from comparative modeling can be used to help understand the details of protein function at an atomic scale, map conservation and mutagenesis data onto a structural framework, and explore detailed functional relationships between protein with similar folds or active sites.

At the other end of the prediction resolution spectrum, *de novo* structure prediction and fold recognition methods produce models of lower resolution than comparative models (see Chapter 10). These models can be used to assign putative functions to proteins for which little is known [15]. At the most basic level we can use structural similarities between a predicted structure and known structures to explore possible distant evolutionary relationships between query proteins of unknown function and other well-studied proteins for which structures have been experimentally determined. A query protein is likely to share some functional aspects with proteins in the PDB that show strong structure–structure matches to a high confidence predicted structure for that protein. This is based on the assumption that detectable structure relationships are conserved across a greater evolutionary distance than are detectable sequence similarities. This assumption is well supported by multiple surveys of the distributions of folds and their related functions in the PDB [48,68,76,83]. The relationship between fold and function, however, is by no means a simple subject, and I refer the reader to several works that discuss this relationship in greater detail [56, 70, 84, 107]. Another way of exploring the functional significance of high confidence predicted structures is to use libraries of three-dimensional (3-D) functional motifs to search for conserved active site or functional motifs on the predicted structures [33, 72, 103]. Both basic methods, fold–fold matching and the use of small 3-D functional motif searches, can in principle be combined to form the basis for deriving functional hypothesis from predicted structure, thereby extending the completeness of genome annotations based only on primary sequence. For more details on how to infer protein function from protein structure, see Chapter 34.

1.3 *De novo* Structure Prediction in a Genome Annotation Context, Synergy with Other Methods

To date, the annotation of protein function in newly sequenced genomes relies on a large array of tools based ultimately on primary sequence analysis [3, 9, 19, 100]. These tools have afforded great progress in genome annotation including large improvements in gene detection, sequence alignment and detection of homologous sequences across genomes as well as the creation of databases of common protein families and primary sequence functional motifs. Comparative modeling methods have been highly successful on many fronts, creating large databases of highly accurate structure predictions for many organisms, but are based on primary sequence matches between PDB and query sequences [87] (see Chapter 10). Primary sequence methods also exist for the prediction of basic local structure qualities (some of these patterns being lower complexity patterns) of sequences such as the location of coiled-coil, transmembrane and disordered regions [52, 80, 99, 104]. Efforts to use *de novo* structure prediction (and/or fold recognition) must employ these sequence-based methods, as these methods provide a solid foundation on which all *de novo* methods discussed herein are reliant (see Figure 1). Any organization of these methods into an annotation pipeline must properly account for the fact that the accuracy/reliability is quite different between sequence and structure-based methods. One approach is to use structure prediction as part of a hierarchy where methods yielding high-confidence results are exhausted prior to computationally expensive and less accurate *de novo* structure prediction and fold recognition ■ [12]■ . I will describe some early results from these approaches/pipelines that include structure prediction, the

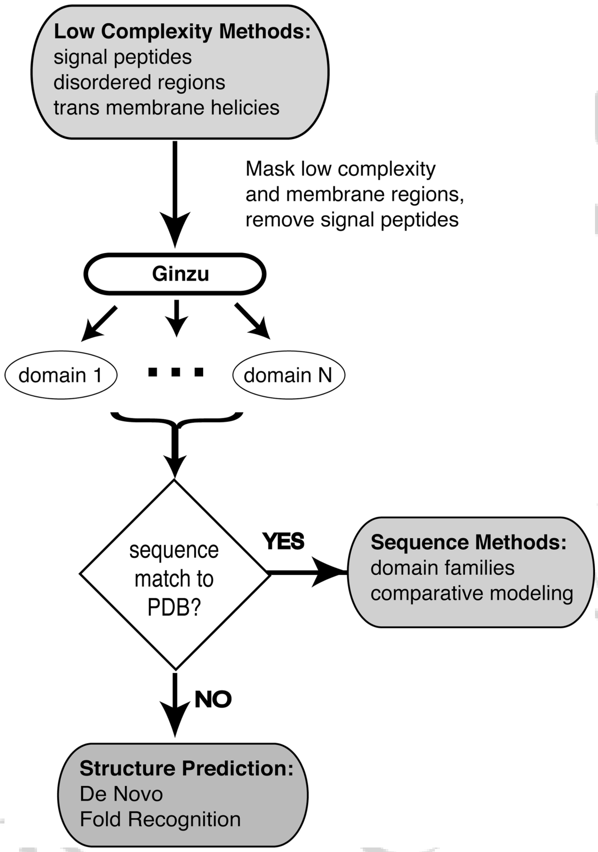

Figure 1 Idealized proteome structure annotation pipeline. Low-complexity regions such as transmembrane helices, signal peptides and disordered regions are masked, and domains dominated by these low-complexity or transmembrane sequence are treated separately. Remaining sequences are parsed to separate regions into structural domains to the degree that such domains are detectable (here, Ginzu is shown as the domain parsing algorithm, see Figure 4). Domains that do not have strong sequence matches to the PDB or other matches to well-annotated domains (Pfam, COG) are forwarded to structure-based methods. The use of structure prediction methods is positioned within this hierarchy of methods to increase comprehension of the resulting annotation without compromising the results obtained by sequence-based methods.

details of these pipelines and the technical and research challenges that remain in applying these pipelines to genome annotation [6,45,86].

The need for methods for predicting transmembrane proteins and understanding membrane–protein interactions is not discussed in this work (see Chapter 9 for this topic), the focus here is instead on soluble domains (including soluble domains excised from proteins containing transmembrane regions). Part of the difficulty in predicting transmembrane protein structure lies in the paucity of membrane protein structures deposited in the PDB

[28, 99]. It is only with access to the PDB, an ideal and comprehensive gold standard, by many criteria, that we can approach the problem of predicting soluble protein structure.

2 Core Features of Current Methods of *De Novo* Structure Prediction

We will now discuss core concepts that are common to multiple successful current *de novo* methods. This review is not intended to be encyclopedic and will invariably fail to mention several methods that are innovative and/or accurate in its attempt to focus on core concepts instead of distinct methodologies. The omission of any specific method should not be interpreted as commentary on the relative accuracy of the omitted method, but is simply due to the scope of this work and the state of rapid development in this field.

2.1 Rosetta *De Novo*

Throughout this work I will use examples of key concepts in *de novo* structure prediction with several examples drawn from the Rosetta *de novo* structure prediction protocol and will thus provide a brief overview of Rosetta before continuing to discuss key elements of the procedure in greater detail [13, 90, 97, 98] (see Figures 2 and 3). Results from the fourth and fifth Critical Assessments of Structure Prediction (CASP4, CASP5 and CASP6; see also Chapter 11) have shown that Rosetta is currently one of the best methods for *de novo* protein structure prediction and distant fold recognition [16, 18, 26, 65]. Rosetta was initially developed as a computer program for *de novo* fold prediction, but has been expanded to include design, docking, experimental determination of structure from partial datasets, protein–protein interaction and protein–DNA interaction prediction [25, 41, 42, 57, 59, 60, 88, 89]. When referring to Rosetta in this work I will be primarily referring to the *de novo* or *ab initio* mode of the Rosetta code base. Early progress in high-resolution structure prediction has been achieved via combinations of low-resolution approaches (for initially searching the conformational landscape) and higher-resolution potentials (where atomic detail and physically derived energy functions are employed). Thus, Rosetta structure prediction is carried out in two phases: (i) a low-resolution phase where overall topology is searched using a statistical scoring function and fragment assembly, and (ii) an atomic-detail refinement phase using rotamers and small backbone angle moves, and a more physically relevant (detailed) scoring function. The algorithms for searching the landscape are Monte-Carlo-type in both phases.

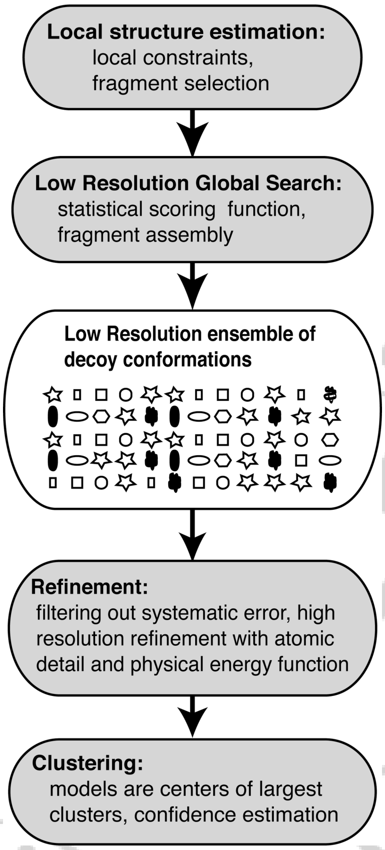
Local structure estimation:
local constraints,
fragment selection
Low Resolution Global Search:
statistical scoring function,
fragment assembly
Low Resolution ensemble of
decoy conformations
Refinement:
filtering out systematic error, high
resolution refinement with atomic
detail and physical energy function
Clustering:
models are centers of largest
clusters, confidence estimation

Figure 2 Schematic outline of Rosetta structure prediction protocol. Single sequences enter at the top of this schematic and confidence-ranked structure predictions are produced by the last/bottom step.

In the first phase, Rosetta *de novo* (Rosetta) uses information from the PDB to estimate the possible conformations for local sequence segments. The procedure first generates libraries of local sequence fragments excised from the PDB on the basis of local sequence similarity (three- and nine-residue matches between the query sequence and a given structure in the PDB). See Figure 1 for a schematic overview of the low-resolution (or fold prediction) phase of the Rosetta method, and see Tables 1 and 2 for a complete description of the Rosetta score. Rosetta fragment generation works well even for sequences that have no homologs in the known sequence databases; the structures in the PDB cover possible local sequence well at the three- and nine-residue length according to the current method. Rosetta then assembles these pre-computed local structure fragments by minimizing a global scoring function that favors hydrophobic burial/packing, strand pairing, compactness and highly proba-

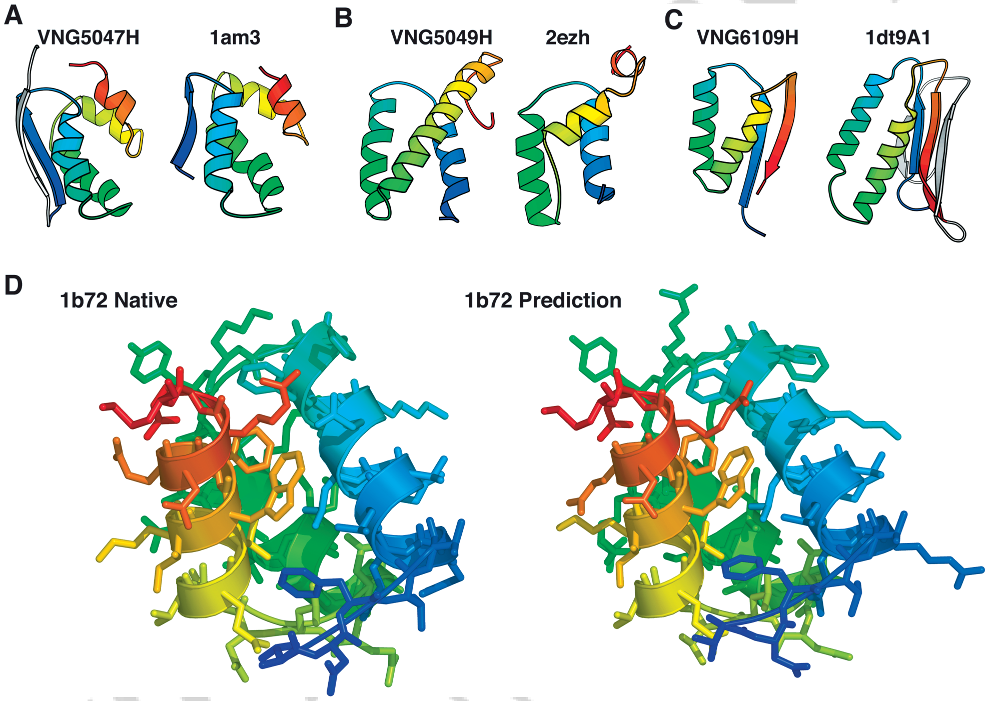

Figure 3 Examples of *de novo* structure predictions generated using Rosetta. (A–C) Examples from our genome-wide prediction of domains of unknown function in *Halobacterium* NRC-1 [12]. In each case the predicted structure is shown next to the correct native. For (A–C) only the backbone ribbons are shown, as these predictions were not refined using the all-atom potential and are examples of the utility of low-resolution prediction in determining function. (D) A recent prediction where high-resolution refinement subsequent to the low-resolution search produced the lowest energy conformation, a prediction of unprecedented accuracy (provided by Phil Bradley) [17].

ble residue pairings. The Rosetta score for this initial low-resolution stage is described in its entirety in Table 1. For the second, refinement, stage centroid representations of amino acid side-chains are replaced with atomic detail (rotamer representations). The scoring function used during this refinement phase includes solvation terms, hydrogen bond terms and other terms with direct physical interpretation. See Table 2 for a full description of the all-atom Rosetta score. Features of the high- and low-resolution phases of the Rosetta method are described below as I discuss key components of *de novo* structure prediction universal to all successful methods.

Using Rosetta generated structure predictions we were able to recapitulate many functional insights not evident from sequence based methods alone [14, 15]. We have reported success in annotating proteins and protein families without links to known structure with Rosetta [8, 14]. Various aspects of this

Table 1 Low-resolution, centroid-based Rosetta scoring function[a]

Name	Description (physical origin)	Functional form	Parameters (values)
env[b]	residue environment (solvation)	$\sum_i -\ln[P(aa_i \| nb_i)]$	i = residue index aa = amino acid type nb = number of neighboring residues[c] (0, 1, 2, …, 30. >30)
pair[b]	residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(aa_i, aa_j \| s_{ij} d_{ij})}{P(aa_i \| s_{ij} d_{ij}) P(aa_j \| s_{ij} d_{ij})}\right]$	i, j = residue indices aa = amino acid type d = centroid–centroid distance (10–12, 7.5–10, 5–7.5, <5 Å) s = sequence separation (>8 residues)
vdw[g]	steric repulsion	$\sum_i \sum_{j>i} \frac{(r_{ij}^2 - d_{ij}^2)^2}{r_{ij}^2}; \quad d_{ij} < r_{ij}$	i, j = residue (or centroid) indices d = interatomic distance r = summed van der Waals radii[h]
rg	radius of gyration (van der Waals attraction; solvation)	$\sqrt{\langle d_{ij}^2 \rangle}$	i, j = residue indices d = distance between residue centroids
cbeta	C_β density (solvation; correction for excluded volume effect introduced by simulation)	$\sum_i \sum_{sh} -\ln\left[\frac{P_{\text{compact}}(nb_{i,sh})}{P_{\text{random}}(nb_{i,sh})}\right]$	i = residue index sh = shell radius (6, 12 Å) nb = number of neighboring residues within shell[f] $P_{compact}$ = probability in compact structures assembled from fragments P_{random} = probability in structures assembled randomly from fragments

overall protocol will be reviewed in greater detail below. We also encourage the reader to refer to several prior works where the Rosetta method is described in its entirety.

2.2 Evaluation of Structure Predictions

In general the most effective methods for predicting structure *de novo* depend on parameters ultimately derived from the PDB. Several methods use the PDB directly to estimate local sequence and even explicitly use fragments of local sequence from the PDB to build global conformations. These uses of the PDB require that methods be tested using structures not present in the sets

Table 1 continued

Name	Description (physical origin)	Functional form	Parameters (values)
SS[d]	strand pairing (hydrogen bonding)	Scheme A: $\mathrm{SS}_{\phi,\theta}$ + SS_{hb} + SS_{d} Scheme B: $\mathrm{SS}_{-\phi,\theta}$ + SS_{hb} + $\mathrm{SS}_{d\sigma}$ where $SS_{\phi,\theta} = \sum_{m}\sum_{n>m} - \ln[P(\phi_{mn}, \theta_{mn} \| d_{mn}, sp_{mn}, s_{mn})]$ $SS_{hb} = \sum_{m}\sum_{n>m} - \ln[P(hb_{mn}, \| d_{mn}, s_{mn})]$ $SS_{d} = \sum_{m}\sum_{n>m} - \ln[P(d_{mn}, \| s_{mn})]$ $SS_{d\sigma} = \sum_{m}\sum_{n>m} - \ln[P(d_{mn}, \sigma_{mn} \| \rho_{m}, \rho_{n})]$	m, n = strand dimer indices; dimer is two consecutive strand residues $\hat{V}$ = vector between first N and last C atom of dimer $\hat{m}$ = unit vector between $\hat{V}m$ and $\hat{V}n$ midpoints $\hat{x}$ = unit vector along carbon-oxygen bond of first dimer residue $\hat{y}$ = unit vector along oxygen-carbon bond of second dimer residue ϕ, θ = polar angles between $\hat{V}m$ and $\hat{V}n$ (10, 36° bins) $hb = dimertwist,$ $\sum_{k=m,n} 0.5(\|\hat{m} \cdot \hat{x}_k\| + \|\hat{m} \cdot \hat{y}_k\|)$ (<0.33, 0.33–0.66, 0.66–1.0, 1.0–1.33, 1.33–1.6, 1.6–1.8, 1.8–2.0) d = distance between $\hat{V}m$ and $\hat{V}n$ midpoints (<6.5 Å) σ = angle between $\hat{V}m$ and $\hat{M}$ (18° bins) sp = sequence separation between dimer-containing strands (<2, 2–10, >10 residues) s = sequence separation between dimers (>5 or >10) ρ = mean angle between vectors $\hat{m}$, $\hat{x}$ and $\hat{m}, \hat{y}$ (180° bins)
sheet[e]	strand arrangement into sheets	$-\ln[P(n_{\mathrm{sheets}} n_{\mathrm{lone\_strands}} \| n_{\mathrm{strands}})]$	n_{sheets} = number of sheets $n_{\mathrm{lone\_strands}}$ = number of unpaired strands n_{strands} = total number of strands
HS	helix-strand packing	$\sum_n \sum_n - \ln[P(\phi_{mn}, \psi_{mn} \| sp_{mn} d_{mn})]$	m = strand dimer index; dimer is two consecutive strand residues n = helix dimer index; dimer is central two residues of four consecutive helical residues $\hat{V}$ = vector between first N and last C atom of dimer ϕ, θ = polar angles between $\hat{V}m$ and $\hat{V}n$ (36° bins) sp = sequence separation between dimer-containing helix and strand (binned <2, 2–10, >10 residues) d = distance between $\hat{V}m$ and $\hat{V}n$ midpoints (<12 Å)

of protein structures used to train these methods (or present in the sets of structures used to predict local structure fragments). The first such evaluation of structure prediction, CASP (see Chapter 11 for a more detailed description), showed that published estimates of prediction error were smaller than prediction error measured on a set of novel proteins outside the training set (this is not surprising given the difficulties of avoiding overfitting in as complex a data space as protein structure) [64]. Indeed, early experiments showed that no methods for *de novo* structure prediction were effective outside of carefully chosen benchmarks containing only the smallest proteins. Spurred on by these early evaluations the field returned to the drawing board and two years later produced multiple methods with much higher accuracies in the new folds or *de novo* category (CASP3) [73,75,82]. Thus, the CASP experiments proved to be invaluable to the field at that point in the development of the field, provoking a renewed interest in the *de novo* structure prediction and properly realigned interest in techniques according to effectiveness.

Arguably, CASP has the flaw that predictors are allowed to intervene and manually curate their predictions prior to submission to the CASP evaluators. Thus, the results of CASP are a convolution of: (i) the art of prediction (each group's intuition and skill using their tools) and (ii) the relative performance

Footnotes to Table 1:

[a] The individual components in the Rosetta score (the score used by Rosetta during low-resolution/centroid mode *de novo* structure prediction) are given as described originally in Simons [96–98].

[b] Binned function values are linearly interpolated, yielding analytic derivatives.

[c] Neighbors within a 10 Å radius. Residue position defined by C_β coordinates (C_α for glycine).

[d] Interactions between dimers within the same strand are neglected. Favorable interactions are limited to preserve pairwise strand interactions, i.e. dimer m can interact favorably with dimers from at most one strand on each side, with the most favorable dimer interaction ($SS_{\phi s\theta}$ + SS_{hb} + SS_{d}) determining the identity of the interacting strand. $SS_{d\sigma}$ is exempt from the requirement of pairwise strand interactions. SS_{hb} is evaluated only for m, n pairs for which $SS_{\phi,\theta}$ is favorable. $SS_{d\sigma}$ is evaluated only for m, n pairs for which $SS_{\phi,\theta}$g and SS_{hb} are favorable. A bonus is awarded for each favorable dimer interaction for which $|m \; - - n| > 11$ and strand separation is more than eight residues

[e] A sheet is comprised of all strands with dimer pairs less than 5.5 Å apart, allowing each strand having at most one neighboring strand on each side. Discrimination between alternate strand pairings is determined according the most favorable dimer interaction. Probability distributions fitted to $c(n_{\text{strands}}) - 0.9n_{\text{sheets}} - 2.7n_{\text{lone\_strands}}$ where $c(n_{\text{strands}})$ = (0.07, 0.41, 0.43, 0.60, 0.61, 0.85, 0.86, 1.12).

[f] Residue position defined by C_β coordinates (C_α for glycine).

[g] Not evaluated for atom (centroid) pairs whose interatomic distance depends on the torsion angles of a single residue.

[h] Radii determined from (i) 25th closest distance seen for atom pair in pdbselect25 structures, (ii) the fifth closest distance observed in X-ray structures with better than 1.3-Å resolution and less than 40% sequence identity or (iii) X-ray structures of less than 2 Å resolution, excluding i, $i + 1$ contacts (centroid radii only).

of the core methods (the performance of each method in an automatic setting). Although this convolution reflects the reality when workers aim to predict proteins of high interest, such as proteins involved in a specific function or proteins critical to a given disease or process being experimentally studied, it does not reflect the demands placed on a method when trying to predict whole genomes, where the shear number of predictions does not allow for much manual intervention. Several additional tests similar to CASP (in that they are blind tests of structure prediction) have been organized in response to the concerns of many that it is important to remove the human aspects of CASP. The Critical Assessment of Fully Automatic Structure Prediction (CAFASP) is an experiment running parallel with CASP that aims to test fully automated methods' performance on CASP targets, mainly testing servers instead of groups [35, 36]. Several groups have also raised concerns that there are problems associated with the small numbers of proteins tested in each CASP experiment, and thus EVA and LiveBench were organized to test methods using larger numbers of proteins [20, 92, 94]. Both use proteins that have structures that are unknown to the participating prediction groups, but that have been recently submitted to the PDB and are not open to the public at the time their sequences are released to those participating in LiveBench or EVA. The participating groups then have the time it takes for the new PDB entries to be validated to predict the structures. Although groups with amazing computer-hacking skills could in principle access this information, these efforts effectively create a CAFASP equivalent for a larger number of proteins.

All four of these tests of prediction methods, as well as benchmarks carried out by authors of any methods in question, are valuable ways of judging the performance of *de novo* methods. The methods, and elements of methods, I describe herein are generally accepted to be the best performers by the five above measures (four blind tests and author benchmarks). I will not focus on the details of the CASP, CAFASP, EVA and LiveBench methods, as they are described in detail elsewhere (see Chapters ■–■) and instead attempt to focus on common elements of top performing methods.

2.3 Domain Prediction is Key

As the size of a protein increases, so to does the size of the conformational space associated with that protein. Thus, *de novo* methods, which must sample this space, have run times that increase dramatically with sequence length. Current *de novo* methods are limited to proteins and protein domains less than 150 amino acids in length (with Rosetta the limit is around 150 residues for α/β proteins, 80 for β-folds and more than 150 residues for α-only-folds). This limit means that roughly half of the protein domains seen so far in the

Table 2 All-atom Rosetta scoring function: the components of the all-atom score (centroids are expanded using a rotamer description of side-chains) [31, 44, 58, 62, 77, 105]

Name	Description	Functional form	Parameters, variables	References
rama	Ramachandran torsion preferences	$\sum_i - \ln[P(\phi_i, \psi_i \vert aa_i ss_i)]$	i = residue index ϕ, ψ = backbone torsion angles (10°, 36° bins) aa = amino acid type ss = secondary structure type[a]	Bowers, 2000 [■]
LJ[c]	Lennard–Jones interactions	$\sum_i \sum_{j>i} \begin{cases} \left\lfloor \left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6} \right\rfloor e_{ij} \\ \text{if} \quad \frac{d_{ij}}{r_{ij}} > 0.6 \\ \left[-8759.2\left(\frac{d_{ij}}{r_{ij}}\right) + 5672.0\right] e_{ij}, \\ \text{else} \end{cases}$	i,j = residue indices d = interatomic distance e = geometric mean of atom well depths[d] r = summed van der Waals radii[e]	Kuhlman, 2000 [■]
hb[f]	hydrogen bonding	$\sum_i \sum_j (- \ln[p(d_{ij} \vert h_j ss_{ij})]$ $- \ln[P(\cos\theta_{ij} \vert d_{ij} h_j ss_{ij})]$ $- \ln[P(\cos\theta_{ij} \vert d_{ij} h_j ss_{ij})]$ $- \ln[P(\cos\psi_{ij} \vert d_{ij} h_j ss_{ij})]$	i = donor residue index j = acceptor residue index d = acceptor-proton interatomic distance h = hybridization (sp^2, sp^3) ss = secondary structure type[g] θ = proton–acceptor–acceptor base bond angle ψ = donor–proton–acceptor bond angle	Kortemme, 2003 [■]
solv	solvation	$\sum_i \left\lfloor \Delta G_i^{\text{ref}} - \sum_j \left(\frac{2\Delta G_i^{\text{free}}}{4\pi^{3/2}\lambda_i r_{ij}^2} e^{-d_{ij}^2} V_j \right.\right.$ $\left.\left. + \frac{2\Delta G_i^{\text{freee}}}{4\pi^{3/2}\lambda_j r_{ij}^2} e^{-d_{ij}^2} V_i \right) \right\rfloor$	i, j = atom indices d = distance between atoms r = summed van der Waals radii[e] λ = correlation length[h] V = atomic volume[h] ΔG^{ref}, ΔG^{free} = energy of a fully solvated atom[h]	Lazaridis, 1999 [■]
pair	residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} - \ln \left\lfloor \frac{P(aa_i, aa_j) \vert d_{ij})}{P(aa_i \vert d_{ij}) P(aa_i \vert d_{ij})} \right\rfloor$	i,j = residue indices aa = amino acid type d = distance between residues[i]	Kuhlman, 2000 [■]
dun	rotamer self energy	$\sum_i - \ln \left\lfloor \frac{P(\text{rot}_i \vert \phi_i \psi_i) P(aa_i \vert \phi_i, \psi_i)}{P(aa_i)} \right\rfloor$	i,j = residue indices rot = Dunbrack backbone-dependent rotamer aa = amino acid type ϕ, ψ = backbone torsion angles	Dunbrack, 1997 [■]
ref	unfolded state reference energy	$\sum_{aa} n_{aa}$	aa = amino acid type n = number of residues	Kuhlman, 2000 [■]

PDB are within the size limit of *de novo* structure prediction. Two approaches to circumventing this size limitation are: (i) increasing the size range of *de novo* structure prediction and (ii) dividing proteins into domains prior to attempting to predict structure. Dividing query sequences into their smallest component domains prior to folding is one straightforward way to dramatically increase the reach of *de novo* structure prediction. For many proteins domain divisions can be easily found (as would be the case for a protein where one domain was unknown and one domain was a member of a well-known protein family) while several domains remain beyond our ability to correctly detect them. The determination of domain family membership and domain boundaries for multi-domain proteins is a vital first step in annotating proteins on the basis of primary sequence and has ramifications for several aspects of protein sequence annotation; multiple works describe methods for detecting such boundaries. In short, most protein domain parsing methods rely on hierarchically searching for domains in a query sequence with a collection of primary sequence methods, domain library searches and matches to structural domains in the PDB [26, 55, 66].

Some notable works use coarse-grained structural simulations/predictions coupled with methods for assigning structural domain boundaries to 3-D structures to detect protein domains from sequence. The guiding principle behind this approach is that very low-resolution predictions will pick up overall patterns of the polypeptide packing into distinct structural domains. Another recent work attempted to use local sequence signals to detect structure domain boundaries under the assumption that there would be detectable differences in local sequence propensities at domain boundaries [37]. As of yet these

Footnotes to Table 2:

[a] All binned function values are linearly interpolated, yielding analytic derivatives, except as noted.
[b] Three-state secondary structure type as assigned by DSSP.
[c] Not evaluated for atom pairs whose interatomic distance depends on the torsion angles of a single residue.
[d] Well depths taken from CHARMm19 parameter set (Neria 1996 [■]).
[e] Radii determined from fitting atom distances in protein X-ray structures to the 6–12 Lennard–Jones potential using CHARMm19 well depths.
[f] Evaluated only for donor acceptor pairs for which $1.4 \leq d \leq 3.0$ and $90^\circ \leq \psi, \theta \leq 180^\circ$. Side-chain hydrogen bonds in involving atoms forming main-chain hydrogen bonds are not evaluated. Individual probability distributions are fitted to eighth-order probability distributions and analytically differentiated.
[g] Secondary structure types for hydrogen bonds are assigned as helical ($j - - i = 4$, main-chain), strand: ($|j - - i| > 4$, main-chain) or other.
[h] Values taken from Lazaridis and Karplus [62].
[i] Residue position defined by C_β coordinates (C_α of glycine).
* Also described in Rohl 2005 [■].

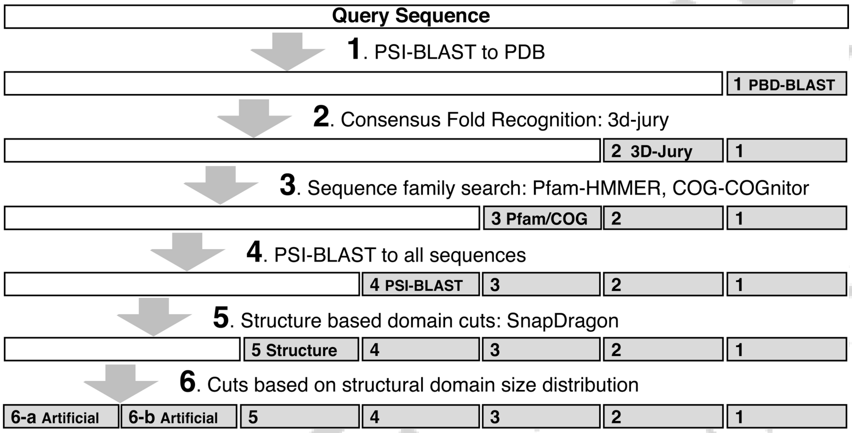
Query Sequence
1. PSI-BLAST to PDB
1 PBD-BLAST
2. Consensus Fold Recognition: 3d-jury
2 3D-Jury
1
3. Sequence family search: Pfam-HMMER, COG-COGnitor
3 Pfam/COG
2
1
4. PSI-BLAST to all sequences
4 PSI-BLAST
3
2
1
5. Structure based domain cuts: SnapDragon
5 Structure
4
3
2
1
6. Cuts based on structural domain size distribution
6-a Artificial
6-b Artificial
5
4
3
2
1

Figure 4 Schematic outline of an ideal hierarchical approach to domain parsing. Methods with higher reliability are used first, with sequence matches to the PDB being the highest-quality information. As higher-reliability/interpretability methods are exhausted, noisier methods are used (such as parsing multiple sequence alignments, step 4, and guessing domain boundaries based on the distribution of domain sizes in the PDB). Sequence regions hit by higher confidence methods (represented as gray rectangles) are masked and the remaining sequence (represented by white rectangles) are forwarded onto the remaining methods. Steps 1–4 and 6 are currently implemented in the Ginzu program; step 5 (adding sequence homolog independent methods such as structure-based domain parsing from sequence to the procedure) represent future work. Although we recognize domains in this schematic from left to right this direction is merely schematic, and Ginzu recognizes and parses domains in a fully general (discontinuous, depending on where the strong hits are at any given level) manner.

methods have unacceptably high error rates and are far too computationally demanding for use in genome wide predictions (David Kim, personal communication) [38]. In spite of the limitations mentioned above, these methods (that are not dependent on detecting sequence homologs for a given query sequence) are attractive for proteins that have no detectable homologs or matches to protein domain families and future work on this front could increase the number of proteins within reach of *de novo* methods considerably. It is likely that a method which successfully combines these coarse structure-based methods with existing sequence-based methods into a hierarchically organized domain detection program (e.g. Ginzu) will eventually outperform any existing method at domain parsing and greatly increase the accuracy of downstream structure prediction. Figure 4 shows a schematic domain detection program (this schematic is implemented as the program Ginzu).

2.4 Local Structure Prediction and Reduced Complexity Models are Central to Current *De Novo* Methods

Several methods for reducing the combinatorial complexity of the protein folding problem have been employed including lattice models (confining possible special coordinates to a predefined 3-D grid) and several discrete-state off-lattice models (e.g. reducing degrees of freedom along the backbone to a set of discrete angles). For a more exhaustive description of these methods and their reduced-complexity move sets I refer the reader to earlier reviews of *de novo* structure prediction methods [11, 27].

Instead, I will focus on the use of local structure information to constrain global structure prediction simulations to only conformations consistent with local structure prediction. Local sequences excised from protein structures often have stable structures in the absence of their global contacts, demonstrating that local sequences can have a strong, sequence-dependent, structural bias towards one or more well defined structures [10, 24, 69, 74, 106]. This experimental observation is a result of the fact that the polypeptide chain is heavily constrained by local structure bias in a sequence dependent manner. The strength of this local, sequence-dependent, structure bias can vary from strong (a local sequence that exhibits a single well defined local structure) to weak (local sequences that are disordered or completely determined by their global environment) with most protein sequences falling into some intermediate regime (local sequences that fold into multiple well-defined local structures depending on their global environment) [21, 46]. Prediction methods that accurately predict the type, strength and possible multiplicity of local structure bias for any given query sequence segment drastically reduce the size of the available conformational landscape. Using either fragment substitution (assembling fragments of local structure) as a move set or local structure constraints derived from predicted local structure also has the advantage that the subsequent global search is limited to protein-like regions of the conformational landscape (helices, correct chirality of secondary strand packing, strands and sheets with correct twist, etc.).

There are two main ways to use local structure prediction as an overriding/hard constraint on the global search: (i) using fragments to build up global structures (local structure defining the moveset) and (ii) using local structure as a hard constraint (local structure heavily modifying the objective function).

Rosetta explicitly uses fragments of three and nine residues of local structure to build global structures via fragment assembly. Prior to a Rosetta simulation a library of local structure fragments is generated such that several fragments (25–200) of different local structure are pre-computed for every possible three- and nine-residue window along the query. The simulation (the search for low-energy conformations given the Rosetta scoring function) consists primarily of randomly selecting three- and nine-residue windows along the query and replacing torsion angles at that three- or nine-residue window with torsion angles taken from a different fragment for that position. These fragments are pulled from a nonredundant version of the PDB on the basis of local sequence similarity to the query sequence [97]. This work was inspired by careful studies of the relationship between local sequence and local structure [46], that demonstrated that this relationship was highly variable on a sequence-specific basis and that there is a great deal of sequence-specific local structure that could be recognized even in the absence of global homology. The selection of fragments of local structure on the basis of local sequence matches dramatically reduces the size of the accessible conformational landscape. In practice we see that, as desired, for some local sequence segments there is a strong bias towards a single local structure in the computed local structure fragments, while other local sequences exhibit a wide range of local conformations in the fragment library. Using fragment substitution as a moveset to optimize Rosetta's objective function has one major drawback: as the structure collapses (forms many contacts favorable according to the energy function) late in the simulation the acceptance rate of fragment moves becomes unworkably small. This is due to the fact that the substitution of six or 18 backbone dihedral angles creates large perturbations to the Cartesian coordinates of parts of the protein distant along the polypeptide chain. The likelihood that such large perturbations cause steric clashes and break energetically favorable contacts late in a given simulation is exceedingly large. To recover effective minimization of the Rosetta score after initial collapse several additional move types have been added to the Rosetta moveset. The simplest move type consists of small angle moves (within populated regions of the Ramachandran map). Additional moves, descriptively named "chuck", "wobble" and "gunn" moves, aim to perform fragment insertions that have small effects far from the insertion. These additional move types are also critical to the modeling of loops in homology modeling and are described in detail elsewhere ■ [89]■ .

The TASSER method smoothly combines fragments of aligned protein structure (from threading runs) with regions of unaligned proteins (represented on a lattice for computational efficiency) to effectively scale between the fold recognition and *de novo* regime [108]. Other notable uses of local structure fragments include the use of I-sites to select fragments that are then fed to Rosetta as described by Bystroff and Shao [22]. I-sites is a hidden Markov model (HMM) method designed to detect strong relationships between sequence and structure as defined by a library of local structure–sequence relationships. One potential advantage of this method is that the I-sites method is not constrained to fragments of a fixed length (Rosetta is constrained to three- and nine-length fragments) [23]. Thus larger patterns of local structure bias are expected to be detected better by this method. Karplus and coworkers also use a similar approach to detecting fragments of local structure (a two-stage HMM) as part of their *de novo* method [53]. These methods have the primary advantage of better performance when local sequence–structure bias is high (e.g. when local structure is strongly and/or uniquely determined by sequence).

2.5 Clustering as a Heuristic Approach to Approximating Entropic Determinants of Protein Folding

Several protein structure prediction methods are effectively two-step procedures involving the generation of large ensembles of conformations (each being the result of a minimization or simulation) followed by the clustering of the generated ensemble to produce one or more cluster centers that are taken to be the predicted models. Regardless of how one justifies the use of clustering as a means of selecting small numbers of predictions or models from ensembles of decoys conformations, the justification is indirectly supported by the efficacy of the procedure and the resultant observation that clustering has become a central, seemingly required, feature of successful *de novo* prediction methods. Starting with CASP3 the field has witnessed a proliferation of clustering methods as post-simulation processing steps in protein structure prediction methods [14, 51, 96, 108].

Prediction of protein structure *de novo* using Rosetta relies heavily on a final clustering stage. In the first step a large ensemble of potential protein structures is generated, each conformation being the result of an extensive Monte Carlo search designed to minimize the Rosetta scoring function (see Figure 1). We then apply clustering to find the centers of the largest clusters. These cluster centers are ranked by the size of their originating cluster in the ensemble. The tightness of clustering in the ensemble is also used as a measure of method success (larger tighter clusters indicate a higher probability that the method produced correct fold predictions for a given protein). Each Rosetta simulation/Monte Carlo run can be thought of as a fast quench starting from a random point on the conformational landscape (defined by the local structure estimation/fragments). Many of these fast quenches (individual simulations) results in incorrect conformations that score nearly as well as any correct conformations generated in the full ensemble of decoy conformations, as judged by the Rosetta score (a number of other potentials tested also lack discriminative power at this stage). This lack of discrimination by *de novo* scoring functions is partially the result of inaccuracies in the scoring function, limitations in our ability to search the landscape and the fact that entropic terms are a major contributor to the free energy of folding. In any case, this lack of discrimination is mitigated by a final clustering step and it has been shown that the centers of the largest clusters in a clustered Rosetta decoy ensemble are in most cases the conformations closest to native. The ubiquitous use of clustering can be justified in several ways: clustering can be thought of as (i) a heuristic way to approximate the entropy of a given conformation given the full ensemble of decoy conformations generated for a given protein, (ii) a signal averaging procedure, averaging out errors in the low-resolution scoring function, or (ii) taking advantage of foldable-protein specific energy landscape features such as broad energy wells that are the result of proteins evolving to be robust to sequence and conformational changes from the native sequence or structure (a mix of sequence and configurational entropy) [95].

An interesting alternative to the strategy of clustering ensembles of results from independent minimizations is the use of replica exchange methods. Replica exchange methods employ large numbers of simulations spanning a range of temperatures (defined physically if one uses a physical potential or simply as a constant in the exponent of the Boltzmann equation (see Chapter 11) for probabilistic scoring functions). These independent simulations are carried out in parallel and are allowed to exchange temperatures throughout the run. This simulation strategy ideally allows for a random walk in energy space (and thus better sampling) and can be used to calculate entropic term *post facto*. Replica exchange Monte Carlo has been used successfully in the simulation and prediction of protein structure, and is interesting due to its explicit connection to a physical description of the system and its ability to search low energy states without getting trapped [81].

2.6 Balancing Resolution with Sampling, Prospects for Improved Accuracy and Atomic Detail

Every *de novo* structure prediction procedure must strike a delicate balance between the computational efficiency of the procedure and the level of physical detail used to model protein structure within the procedure. Low-resolution models can be used to predict protein topology/folds and sometimes suggest function [15]. Low-resolution models have also been remarkably successful at predicting features of the folding process such as folding rates and phi values [1, 2]. It is clear, however, that modeling proteins (and possibly bound water and other cofactors) at atomic detail and scoring these higher resolution models with physically derived, detailed potentials is a needed development if higher-resolution structure prediction is to be achieved.

Early progress has focused on the use of low-resolution approaches for initially searching the conformational landscape followed by a refinement step where atomic detail and physical scoring functions are used to select and/or generate higher-resolution structures. For example, several studies have illustrated the usefulness of using *de novo* structure prediction methods as part of a two-stage process in which low-resolution methods are used for fragment assembly and the resulting models are refined using a more physical potential and atomic detail (e.g. rotamers) [31] to represent side-chains [18, 71, 102]. In the first step, Rosetta is used to search the space of possible backbone conformations with all side-chains represented as centroids. This process is well described, and has well-characterized error rates and behavior. High-confidence or low-scoring models are then refined using potentials that account for atomic detail such as hydrogen bonding, van der Waals forces and electrostatics.

One major challenge that faces methods attempting to refine *de novo* methods is that the addition of side-chain degrees of freedom combined with the reduced length scale (reduced radius of convergence; one must get much closer to the correct answer before the scoring function recognizes the conformation as correct) of the potentials employed require the sampling of a much larger space of possible conformations. Thus, one has to correctly determine roughly twice the number of bond angles to a higher tolerance if one hopes to succeed. An illustrative example of the difference in length scale (radius of convergence) between low-resolution methods and high-resolution methods is the scoring of hydrogen bonds. In the low-resolution Rosetta procedure backbone hydrogen bonding is scored indirectly by a term designed to pack strands into sheets under the assumption that correct alignment of strands satisfies hydrogen bonds between backbone atoms along the strand and that intra-helix backbone hydrogen bonds are already well accounted for by the local structure fragments. This low-resolution method first reduces strands to vectors, and then scores strand arrangement (and the correct hydrogen bonding implicit in the relative positions/arrangement of all strand vector pairs) via functions dependent on the angular and distance relationships between the two vectors. Thus, the scoring function is robust to a rather large amount of error in the coordinates of individual electron donors and acceptors participating in backbone hydrogen bonds (as large numbers of residues are reduced to the angle and distance between the two vectors representing a given pair of strands). In the high-resolution, refinement mode of Rosetta an empirical hydrogen bond terms with angle and distance dependence between individual electron donors and acceptors is used [88]. This more-detailed hydrogen bond term has a higher fidelity and a more straightforward connection to the calculation of physically realistic energies (meaningful units), but requires more sampling, as smaller changes in the orientation of the backbone can cause large fluctuations in computed energy.

Another major challenge with high-resolution methods is the difficulty of computing accurate potentials for atomic-detail protein modeling in solvent; with electrostatic and solvation terms being among the most difficult terms to accurately model. Full treatment of the free energy of a protein conformation (with correct treatment of dielectric screening) is complicated by the fact that some waters are detectably bound to the surface of proteins and mediate interactions between residues [34]. Another challenge is the computational cost of full treatment of electrostatic free energy by solving the Poisson–Boltzmann or linearized Poisson–Boltzmann equations for large numbers of conformations. In spite of these difficulties several studies have shown that refinement of *de novo* structures with atomic-detail potentials can increase our ability to select and or generate near native structures [78]. These methods can correctly select near native conformations from these ensembles and improve near native structures, but still rely heavily on the initial low-resolution search to produce an ensemble containing good starting structures [63,71,102]. Some recent examples of high-resolution predictions are quite encouraging and an emerging consensus in the field is that higher resolution *de novo* structure prediction (structure predictions with atomic-detail representations of side-chains) will begin to work if sampling is dramatically increased.

Progress in high-resolution structure prediction will invariably be carried out in parallel with methods including, but not limited to, predicting protein–protein interactions, designing proteins and distilling structures from partially assigned experimental data sets. Indeed, many of the scoring and search strategies that high-resolution *de novo* structure refinement methods employ were initially developed in the context of homology modeling and protein design [61,90].

3 Applying Structure Prediction: *De Novo* Structure Prediction in a Systems Biology Context

Sequence databases are growing rapidly, with new genomes being deposited at a phenomenal pace. A large portion of each of these newly sequenced genomes can be expected to contain proteins that have no detectable homologs or only homologs of unknown function. It can be expected that even with the continued progress of large experimental structural biology efforts there will remain a large number of proteins for which *de novo* structure prediction and distant fold recognition methods are the only options.

3.1 Structure Prediction as a Road to Function

The relationship between protein structure and protein function is discussed in detail in Chapter 33, but will be reviewed briefly here in the context of *de novo* structure prediction. One paradigm for predicting the function of proteins of unknown function in the absence of homologs, sometimes referred to as the "sequence-to-structure-to-structure-to-function" paradigm, is based on the assumption that 3-D structure patterns are conserved across a much greater evolutionary distance than recognizable primary sequence patterns [33]. This assumption is based on the results of several structure–function surveys which show that structure similarities (fold matches between different proteins in the PDB) in the absence of sequence similarities imply some shared function in the majority of cases [48,67,70,84,101]. One protocol for predicting protein function based on this observation is to predict the structure of a query sequence of interest and then use the predicted structure to search for fold or structural similarities between the predicted protein structure and experimentally determined protein structures in the PDB or a nonredundant subset of the PDB [49,76,83,85]. There are several problems associated with deriving functional annotation from fold similarity, e.g. ;old similarities can occur through convergent evolution and thus have no functional implications. Also, aspects of function can change throughout evolution leaving only general function intact across a given fold superfamily [43, 56, 91]. Fold matches between the predicted structures and the PDB are thus treated as sources of putative general functional information, and are functionally interpreted primarily in combination with other methods such as global expression analysis and the predicted protein association network. To circumvent these ambiguities one can (i) use *de novo* structure prediction and/or fold recognition to generate a confidence ranked list of possible structures for proteins or protein domains of unknown function, (ii) search each of the ranked structure predictions against the PDB for fold similarities and possible 3-D motifs, (iii) calculate confidences for the fold predictions and 3-D motif matches, and, finally, (iv) evaluate possible functional roles in the context of the other systems biology data, such as expression analysis, protein interactions, metabolic networks and comparative genomics.

3.2 Initial Application of *De Novo* Structure Prediction

To date there have been few studies using *de novo* structure prediction as a method for genome annotation, due primarily to the computational expense of the calculations and the relative novelty of the methods. These studies have been carried out in combination with a variety of fold recognition and sequence-based methods for gene annotation, and have provided preliminary results that highlight several successes. It is based on these studies that we argue that *de novo* structure prediction is a viable option for exploring genes of unknown function.

The first emergence of *de novo* structure prediction methods for large-scale structure prediction was heavily limited by available computer resources. These studies were essentially pilot studies to evaluate the potential worth of genome-wide *de novo* structure predictions. In one early study workers were limited to generating predictions for 85 proteins in *Mycoplasma genitalium*, producing around 24 correct fold predictions [54]. Another study approached the computational limitation by folding representatives of Pfam protein families of unknown structure and function [14]. Using this method we were able to generate high confidence fold/structure predictions for around 60% of the 510 protein families for which Rosetta predictions were attempted, covering an additional roughly 12% of the sequences available at that time. Subsequent experimental determination for several of these protein families has shown our computed confidence values to be good estimates of our predictive performance, with success rates (rates of correct fold identification) on internal benchmarks and success rates from blind tests (CASP results and recently solved structures) nearly indistinguishable. Alas, the results of this study were not widely used by biologists due partially to the fact that at the time methods for integrating the resultant low-resolution structure predictions with other data types were not in place. The results of these early studies suggested, however, that whole-genome application of *de novo* structure prediction would result in usable annotations if presented to biologists properly, i.e. integrated with other available data types.

3.3 Application on Genome-wide Scale and Examples of Data Integration

Genome-wide measurements of mRNA transcripts, protein concentrations, protein–protein interactions and protein–DNA interactions generate rich sources of data on proteins, both those with known and those with unknown functions [5, 7]. These systems-level measurements seldom suggest a unique function for a given protein of interest, but often suggest their association with or perhaps their direct participation in a previously known cellular process. Investigators using genome-wide experimental techniques are thus routinely generating data for proteins of hitherto unknown function that appear to play pivotal roles in their studies.

The first full-genome application of *de novo* structure prediction was to the genome of *Halobacterium* NRC-1 [12]. This archaeon is an extreme halophile that thrives in saturated brine environments such as the Dead Sea and solar salterns. It offers a versatile and easily assayed system with several well-coordinated physiologies that are necessary for survival in its harsh environment. The completely sequenced genome of *Halobacterium* NRC-1 (containing around 2600 genes) has provided insights into many of its physiological capabilities; however, nearly half of all genes encoded in the halobacterial genome had no known function prior to our re-annotation [29, 30, 79, 93]. A multi-institutional effort is currently underway to study the genome-wide response of *Halobacterium* NRC-1 to its environment, elevating the need for applying improved methods for annotating proteins of unknown function found in the *Halobacterium* NRC-1 genome. Rosetta *de novo* structure prediction was used to predict 3-D structures for 1185 proteins and protein domains (less than 150 residues in length) found in *Halobacterium* NRC-1. Predicted structures were searched against the PDB to identify fold matches [85] and were analyzed in the context of a predicted association network composed of several sources of functional associations, such as predicted protein interactions, predicted operons, phylogenetic profile similarity and domain fusion. This annotation pipeline was also applied to the recently sequenced genome of *Haloarcula marismortui* with similar rates of correct fold identification.

An application of *de novo* structure prediction to yeast has also been described. This study focused on the application and integration of several methods (ranging from experimental methods to *de novo* structure prediction) to 100 essential open reading frames (ORFs) in yeast [47]. For these 100 proteins the group applied affinity purification followed by mass spectrometry (to detect protein binding partners), two-hybrid analysis, florescence microscopy (to localize proteins) and *de novo* structure prediction (Ginzu to separate domains [26, 55] and Rosetta to build structures for domains of unknown function). Due to the cost of experiments and the computational cost of Rosetta *de novo* structure prediction, the group was initially able to prototype the method on just these 100 proteins. Function was assigned to 48 of the proteins (as defined by assignment to Gene Ontology categories). In total, 77 of the 100 proteins were annotated (had confident hits) by on of the methods employed. Given that the starting set represented a difficult set of ORFs of no known function this represents a significant milestone. Scaling this sort of approach up to whole genomes (including large eukaryotic genomes) is still a significant challenge. A grid computing solution (Section 3.4) is currently being employed to complete this study (fold the remaining ORFs in the genome) and, due to the wide use of yeast as a model organism, we can expect this complete resource to be a major step in crossing the social and technical barrier that has so far prevented the wide application of *de novo* structure prediction to biology. A similar approach has also been applied to the Y chromosome of *Homo sapiens* [40]. By integrating fold recognition with *de novo* structure prediction folds were assigned to around 42 of the 60 recognized domains examined (these 60 domains originated from the 27 proteins thought to be encoded on this chromosome at the time of the study). In both of these application, yeast and human, careful thought was put into reducing the set of proteins examined and scaling-up *de novo* structure prediction remains a critical bottleneck (the introduction of all-atom or high-resolution refinement of these predicted structures will only exacerbate this critical need for computing).

3.4 Scaling-up *De Novo* Structure Prediction: Rosetta on the World Community Grid

There are several strategies one can use to limit the number of protein domains for which computationally expensive *de novo* structure prediction needs to be carried out, allowing for the calculation of useful *de novo* structure predictions for only the most relevant subsets of larger genomes, as discussed above. In spite of these strategies, finding the required compute resources has been a constant challenge for the application of *de novo* structure prediction to functional annotation and has limited the application of the method. To circumvent this problem we are currently applying a grid, distributed computing, solution to folding over 100 000 domains with the full Rosetta *de novo* structure prediction protocol (www.worldcommunitygrid.org). These domains were chosen by applying Ginzu [26,55] to over 60 complete genomes as well as several other appropriate sequences in public sequence databases. The results will be integrated with data types that are appropriate/available for a given organism in collaboration with several other groups [12,47]. This work is ongoing in collaboration with David Baker, Lars Malmstroem (University of Washington) Rick Alther, Bill Boverman and Viktors Berstis (IBM), and United Devices (Austin, TX). Currently ■ (11:10 AM, pacific coast time, 14 April 2005)■, there are over 1 million volunteers (people who have downloaded the client to run grid-Rosetta) comprising a virtual grid of over 3 million devices. Interested parties wishing to participate (donate idle CPU time on your desktop computer to this project) can download the grid-enabled Rosetta client at www.worldcommunitygrid.org. This amount of computational power will enable us to remove the barrier represented by the computational cost of *de novo* methods.

4 Future Directions

4.1 Structure Prediction and Systems Biology: Data Integration

Even with dramatically improved accuracy we still face challenges due to the ambiguities of the relationship between fold and function seen for many fold families (indeed, even close sequence homology is not always trivial to interpret as functional similarity, see also Chapter 30). Thus, the full potential of *de novo* structure prediction in a systems biology context can only be realized if structure predictions are integrated into larger analysis, and subsequently made accessible to biologists through better data integration, analysis and visualization tools. One clear example of this is provided by the bacterial transcription factors, for which even strong sequence similarity can imply several possible functions and system-wide information is required to determine a meaningful function (the target of a given transcription factor).

4.2 Need for Improved Accuracy and Extending the Reach of *De Novo* Methods

Although I have argued that data integration is as critical a bottleneck as any other and that there are current applications of *de novo* structure prediction, it is clear that improved accuracy is also essential for progress in the field and for the acceptance of *de novo* structure methods by the end users of whole-genome annotations. There is still a significant amount of error in predictions generated using current structure prediction and domain parsing methods. Extending the size limit of protein folding methods is a promising area of active research as is the development of higher-resolution refinement methods. *De novo* structure prediction requires large amounts of CPU time compared to sequence-based and fold recognition methods (although the use of distributed computing and Moore's law continue to make this less of a bottleneck). Integrating *de novo* predictions with orthogonal sources of general and putative functional information, both experimental and computational, will likely facilitate the annotation of significant portions of the protein sequences resulting from ongoing sequencing efforts, as well as proteins in currently sequenced genomes.

Acknowledgments

I would like to thank Viktors Berstis, Rick Alther and Bill Boverman of IBM for their critical work on the World Community Grid, and Phil Bradley, Lars Malmstroem, David Kim, Dylan Chivian and David Baker for myriad comments, opinions, structure prediction, preprints etc.

References

1 Alm, E. and D. Baker. 1999. Matching theory and experiment in protein folding. Curr. Opin. Struct. Biol. **9**: 189–96.

2 Alm, E. and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc. Natl Acad. Sci. USA **96**: 11305–10.

3 Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–402.

4 Baker, D. and A. Sali. 2001. Protein structure prediction and structural genomics. Science **294**: 93–6.

5 Baliga N. S., S. J. Bjork, R. Bonneau, M. Pan, C. Iloanusi , M. C. H. Kottemann, L. Hood and J. DiRuggiero. 2004. Systems level insights into the stress response to UV radiation in the halophilic archaeon *Halobacterium* NRC-1. Genome Res. **14**: 1025–35.

6 Baliga, N. S., R. Bonneau, M. T. Facciotti, et al. 2004. Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. Genome Res. **14**: 2221–34.

7 Baliga, N. S., M. Pan, Y. A. Goo, et al. 2002. Coordinate regulation of energy transduction modules in *Halobacterium* sp. analyzed by a global systems approach. Proc. Natl Acad. Sci. USA **99**: 14913–8.

8 Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe and E. L. Sonnhammer. 2000. The Pfam protein families database. Nucleic Acids Res. **28**: 263–6.

9 Bateman, A., L. Coin, R. Durbin, et al. 2004. The Pfam protein families database. Nucleic Acids Res. **32**: D138–41.

10 Blanco, F. J., G. Rivas and L. Serrano. 1994. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. Nat. Struct. Biol. **1**: 584–90.

11 Bonneau, R. and D. Baker. 2001. *Ab initio* protein structure prediction: progress and prospects. Annu. Rev. Biophys. Biomol. Struct. **30**: 173–89.

12 Bonneau, R., N. S. Baliga, E. W. Deutsch, P. Shannon and L. Hood. 2004. Comprehensive *de novo* structure prediction in a systems-biology context for the archaea *Halobacterium* sp. NRC-1. Genome Biol. **5**: R52.

13 Bonneau, R., C. E. Strauss and D. Baker. 2001. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins **43**: 1–11.

14 Bonneau, R., C. E. Strauss, C. A. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson and D. Baker. 2002. *De novo* prediction of three-dimensional structures for major protein families. J. Mol. Biol. **322**: 65–78.

15 Bonneau, R., J. Tsai, I. Ruczinski and D. Baker. 2001. Functional inferences from blind *ab initio* protein structure predictions. J. Struct. Biol. **134**: 186–90.

16 Bonneau, R., J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss and D. Baker. 2001. Rosetta in CASP4: progress in *ab initio* protein structure prediction. Proteins **Suppl. 5**: 119–26.

17 Bradley, P., K. M. S. Misura and D. Baker. 2006. Toward high-resolution *de novo* structure prediction for small proteins. Science **309**: 1868–71.

18 Bradley, P., D. Chivian, J. Meiler, et al. 2003. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins **53** (**Suppl. 6**): 457–68.

19 Brenner, S. E., C. Chothia and T. J. Hubbard. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl Acad. Sci. USA **95**: 6073–8.

20 Bujnicki, J. M., A. Elofsson, D. Fischer and L. Rychlewski. 2001. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci. **10**: 352–61.

21 Bystroff, C. and D. Baker. 1998. Prediction of local structure in proteins using a library of sequence–structure motifs. J. Mol. Biol. **281**: 565–77.

22 Bystroff, C. and Y. Shao. 2002. Fully automated *ab initio* protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinformatics **18** (**Suppl. 1**): S54–61.

23 Bystroff, C., V. Thorsson and D. Baker. 2000. HMMSTR: a hidden Markov model for local sequence–structure correlations in proteins. J. Mol. Biol. **301**: 173–90.

24 Callihan, D. E. and T. M. Logan. 1999. Conformations of peptide fragments from the FK506 binding protein: comparison with the native and urea-unfolded states. J. Mol. Biol. **285**: 2161–75.

25 Chevalier, B. S., T. Kortemme, M. S. Chadsey, D. Baker, R. J. Monnat and B. L. Stoddard. 2002. Design, activity, and structure of a highly specific artificial endonuclease. Mol. Cells **10**: 895–905.

26 Chivian, D., D. E. Kim, L. Malmstrom, et al. 2003. Automated prediction of CASP-5 structures using the

Robetta server. Proteins **53** (**Suppl. 6**): 524–33.

27 Chivian, D., T. Robertson, R. Bonneau and D. Baker. 2003. *Ab initio* methods. Methods Biochem. Anal. **44**: 547–57.

28 Deshpande, N., K. J. Addess, W. F. Bluhm, et al. 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res. **33**: D233–7.

29 Devos, D. and A. Valencia. 2001. Intrinsic errors in genome annotation. Trends Genet. **17**: 429–31.

30 Devos, D. and A. Valencia. 2000. Practical limits of function prediction. Proteins **41**: 98–107.

31 Dunbrack, R. L., Jr. and F. E. Cohen. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. **6**: 1661–81.

32 Fetrow, J. S., A. Giammona, A. Kolinski and J. Skolnick. 2002. The protein folding problem: a biophysical enigma. Curr. Pharm. Biotechnol. **3**: 329–47.

33 Fetrow, J. S. and J. Skolnick. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J. Mol. Biol. **281**: 949–68.

34 Finney, J. L. 1977. The organization and function of water in protein crystals. Philos. Trans. R. Soc. Lond. B **278**: 3–32.

35 Fischer, D., C. Barret, K. Bryson, et al. 1999. CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins **Suppl. ■**: 209–17.

36 Fischer, D., L. Rychlewski, R. L. Dunbrack, Jr., A. R. Ortiz and A. Elofsson. 2003. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins **53** (**Suppl. 6**): 503–16.

37 Galzitskaya, O. V. and B. S. Melnik. 2003. Prediction of protein domain boundaries from sequence alone. Protein Sci. **12**: 696–701.

38 George, R. A. and J. Heringa. 2002. SnapDRAGON: a method to delineate protein structural domains from sequence data. J. Mol. Biol. **316**: 839–51.

39 Ginalski, K., N. V. Grishin, A. Godzik and L. Rychlewski. 2005. Practical lessons from protein structure prediction. Nucleic Acids Res. **33**: 1874–91.

40 Ginalski, K., L. Rychlewski, D. Baker and N. V. Grishin. 2004. Protein structure prediction for the male-specific region of the human Y chromosome. Proc. Natl Acad. Sci. USA **101**: 2305–10.

41 Gray, J. J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl and D. Baker. 2003. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. **331**: 281–99.

42 Gray, J. J., S. E. Moughon, T. Kortemme, O. Schueler-Furman, K. M. Misura, A. V. Morozov and D. Baker. 2003. Protein–protein docking predictions for the CAPRI experiment. Proteins **52**: 118–22.

43 Grishin, N. V. 2001. Fold change in evolution of protein structures. J. Struct. Biol. **134**: 167–85.

44 Gunn, J. R. 1997. Sampling protein conformations using segment libraries and a genetic algorithm. J. Chem. Phys. **106**: 4270.

45 Haas, B. J., J. R. Wortman, C. M. Ronning, et al. 2005. Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. BMC Biol. **3**: 7.

46 Han, K. F., C. Bystroff and D. Baker. 1997. Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. Protein Sci. **6**: 1587–90.

47 Hazbun, T. R., L. Malmstrom, S. Anderson, et al. 2003. Assigning function to yeast proteins by integration of technologies. Mol. Cells **12**: 1353–65.

48 Holm, L. and C. Sander. 1997. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res. **25**: 231–4.

49 Holm, L. and C. Sander. 1993. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233**: 123–38.

50 Huang, E. S., R. Samudrala and B. H. Park. 2000. Scoring functions for *ab initio* protein structure prediction. Methods Mol. Biol. **143**: 223–45.

51 Hung, L. H. and R. Samudrala. 2003. PROTINFO: secondary and tertiary protein structure prediction. Nucleic Acids Res. **31**: 3296–9.

52 Jones, D. T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**: 195–202.

53 Karplus, K., R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans and R. Hughey. 2003. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins **53** (**Suppl. 6**): 491–6.

54 Kihara, D., Y. Zhang, H. Lu, A. Kolinski and J. Skolnick. 2002. *Ab initio* protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. Proc. Natl Acad. Sci. USA **99**: 5993–8.

55 Kim, D. E., D. Chivian and D. Baker. 2004. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. **32**: W526–31.

56 Kinch, L. N. and N. V. Grishin. 2002. Evolution of protein structures and functions. Curr. Opin. Struct. Biol. **12**: 400–8.

57 Kortemme, T., L. A. Joachimiak, A. N. Bullock, A. D. Schuler, B. L. Stoddard and D. Baker. 2004. Computational redesign of protein–protein interaction specificity. Nat. Struct. Mol. Biol. **11**: 371–9.

58 Kortemme, T., A. V. Morozov and D. Baker. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. J. Mol. Biol. **326**: 1239–59.

59 Kuhlman, B. and D. Baker. 2000. Native protein sequences are close to optimal for their structures. Proc. Natl Acad. Sci. USA **97**: 10383–8.

60 Kuhlman, B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard and D. Baker. 2003. Design of a novel globular protein fold with atomic-level accuracy. Science **302**: 1364–8.

61 Kuhlman, B., J. W. O'Neill, D. E. Kim, K. Y. Zhang and D. Baker. 2002. Accurate computer-based design of a new backbone conformation in the second turn of protein L. J. Mol. Biol. **315**: 471–7.

62 Lazaridis, T. and M. Karplus. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J. Mol. Biol. **288**: 477–87.

63 Lee, M. R., J. Tsai, D. Baker and P. A. Kollman. 2001. Molecular dynamics in the endgame of protein structure prediction. J. Mol. Biol. **313**: 417–30.

64 Lesk, A. M. 1997. CASP2: report on *ab initio* predictions. Proteins **Suppl. ■:** 151–66.

65 Lesk, A. M., L. Lo Conte and T. J. Hubbard. 2001. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins **Suppl. 5**: 98–118.

66 Liu, J. and B. Rost. 2004. CHOP: parsing proteins into structural domains. Nucleic Acids Res. **32**: W569–71.

67 Lo Conte, L., B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin and C. Chothia. 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res. **28**: 257–9.

68 Lo Conte, L., S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin. 2002. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res. **30**: 264–7.

69 Marqusee, S., V. H. Robbins and R. L. Baldwin. 1989. Unusually stable helix formation in short alanine-based peptides. Proc. Natl Acad. Sci. USA **86**: 5286–90.

70 Martin, A. C., C. A. Orengo, E. G. Hutchinson, et al. 1998. Protein folds and functions. Structure **6**: 875–84.

71 Misura, K. M. and D. Baker. 2005. Progress and challenges in high-resolution refinement of protein structure models. Proteins **59**: 15–29.

72 Moodie, S. L., J. B. Mitchell and J. M. Thornton. 1996. Protein recognition of adenylate: an example of a fuzzy recognition template. J. Mol. Biol. **263**: 486–500.

73 Moult, J., T. Hubbard, K. Fidelis and J. T. Pedersen. 1999. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins **Suppl. ■:** 2–6.

74 Munoz, V. and L. Serrano. 1996. Local versus nonlocal interactions in protein folding and stability – an experimentalist's point of view. Fold. Des. **1**: R71–7.

75 Murzin, A. G. 1999. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. Proteins **Suppl. ■:** 88–103.

76 Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**: 536–40.

77 Neria, E., S. Fischer and M. Karplus. 1996. Simulation of activation free energies in molecular systems. J. Chem. Phys. **105**: 1902.

78 Neves-Petersen, M. T. and S. B. Petersen. 2003. Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules – applications in biotechnology. Biotechnol. Annu. Rev. **9**: 315–95.

79 Ng, W. V., S. P. Kennedy, G. G. Mahairas, et al. 2000. Genome sequence of *Halobacterium* species NRC-1. Proc. Natl Acad. Sci. USA **97**: 12176–81.

80 Nielsen, H., S. Brunak and G. von Heijne. 1999. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng. **12**: 3–9.

81 Okamoto, Y. 2004. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. J. Mol. Graph. Model. **22**: 425–39.

82 Orengo, C. A., J. E. Bray, T. Hubbard, L. LoConte and I. Sillitoe. 1999. Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction. Proteins **Suppl. ■ :** 149–70.

83 Orengo, C. A., F. M. Pearl and J. M. Thornton. 2003. The CATH domain structure database. Methods Biochem. Anal. **44**: 249–71.

84 Orengo, C. A., A. E. Todd and J. M. Thornton. 1999. From protein structure to function. Curr. Opin. Struct. Biol. **9**: 374–82.

85 Ortiz, A. R., C. E. Strauss and O. Olmea. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. **11**: 2606–21.

86 Ouzounis, C. A. and P. D. Karp. 2002. The past, present and future of genome-wide re-annotation. Genome Biol. **3**: COMMENT2001.

87 Pieper, U., N. Eswar, A. C. Stuart, V. A. Ilyin and A. Sali. 2002. MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res. **30**: 255–9.

88 Rohl, C. A. 2005. Protein structure estimation from minimal restraints using Rosetta. Methods Enzymol. **394**: 244–60.

89 Rohl, C. A., C. E. Strauss, D. Chivian and D. Baker. 2004. Modeling structurally variable regions in homologous proteins with Rosetta. Proteins **55**: 656–77.

90 Rohl, C. A., C. E. Strauss, K. M. Misura and D. Baker. 2004. Protein structure prediction using Rosetta. Methods Enzymol. **383**: 66–93.

91 Rost, B. 1997. Protein structures sustain evolutionary drift. Fold. Des. **2**: S19–24.

92 Rost, B. and V. A. Eyrich. 2001. EVA: large-scale analysis of secondary structure prediction. Proteins **Suppl. 5**: 192–9.

93 Rost, B. and A. Valencia. 1996. Pitfalls of protein sequence analysis. Curr. Opin. Biotechnol. **7**: 457–61.

94 Rychlewski, L. and D. Fischer. 2005. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci. **14**: 240–5.

95 Shortle, D., K. T. Simons and D. Baker. 1998. Clustering of low-energy conformations near the native structures

of small proteins. Proc. Natl Acad. Sci. USA **95**: 11158–62.

96 Simons, K. T., R. Bonneau, I. Ruczinski and D. Baker. 1999. *Ab initio* protein structure prediction of CASP III targets using ROSETTA. Proteins **Suppl. 3**: 171–6.

97 Simons, K. T., C. Kooperberg, E. Huang and D. Baker. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. **268**: 209–25.

98 Simons, K. T., I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff and D. Baker. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins **34**: 82–95.

99 Sonnhammer, E. L., G. von Heijne and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. ISMB **6**: 175–82.

100 Tatusov, R. L., N. D. Fedorova, J. J. Jackson, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4**: 41.

101 Todd, A. E., C. A. Orengo and J. M. Thornton. 2001. Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. **307**: 1113–43.

102 Tsai, J., R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl and D. Baker. 2003. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins **53**: 76–87.

103 Wallace, A. C., R. A. Laskowski and J. M. Thornton. 1996. Derivation of 3D coordinate templates for searching structural databases: application to Ser–His–Asp catalytic triads in the serine proteinases and lipases. Protein Sci. **5**: 1001–13.

104 Ward, J. J., L. J. McGuffin, K. Bryson, B. F. Buxton and D. T. Jones. 2004. The DISOPRED server for the prediction of protein disorder. Bioinformatics **20**: 2138–9.

105 Wedemeyer, W. J. and D. Baker. 2003. Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. Proteins **53**: 262–72.

106 Yi, Q., C. Bystroff, P. Rajagopal, R. E. Klevit and D. Baker. 1998. Prediction and structural characterization of an independently folding substructure in the src SH3 domain. J. Mol. Biol. **283**: 293–300.

107 Zhang, B., L. Rychlewski, K. Pawlowski, J. S. Fetrow, J. Skolnick and A. Godzik. 1999. From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. Protein Sci. **8**: 1104–15.

108 Zhang, Y. and J. Skolnick. 2004. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys. J. **87**: 2647–55.